リバーシにおける個人を模倣するプレイヤの作成と評価

1180372　二神　勝俊　【 高度プログラミング研究室 】

1　はじめに

本研究では，特定のプレイヤを模倣したプレイヤの作成を行う．将棋における個人を模倣するプレイヤの作成では，山本らによって評価関数の学習において，大規模なデータセットと小規模なデータセットを組み合わせて用いることで，良い結果が得られる事が示されている [1]．他のゲームや手法にも，参考論文の手法が有用であるのかを調査するため，本実験ではオセロを実験の対象とし，ボナンザメソッドに加えて新たに TD 学習を用いたプレイヤを用意した．

2　ゲームプレイヤにおける機械学習の手法

2.1　TD 学習

TD 学習は強化学習の手法の 1 つで，現在の時間 t における状態 s_t と行動 a_t から求められた行動価値 $Q(s_t, a_t)$ と，時間 t-1 における状態 s_{t-1} と行動 a_{t-1} から求められる行動価値 $Q(s_{t-1}, a_{t-1})$ と得られた報酬 r のの差を比較し，その誤差を 0 に近づける手法である．

$$d = r + Q(s_t, s_a) - Q(s_{t-1}, a_{t-1}) \tag{1}$$

$$Q(s_{t-1}, a_{t-1}) \leftarrow Q(s_{t-1}, a_{t-1}) + \alpha * d \tag{2}$$

ここで α は学習率である．本実験では，報酬 r を終局状態の時の石の差とした．

2.2　ボナンザメソッド

S を盤面，s をその 1 手進めた盤面とする．f は盤面とその特徴ベクトルにより盤面の評価値を返す関数である．i=0 の手を棋譜データとすると，$l(S, v)$ は棋譜中の手と，他の指し手の評価の違いを表現する関数である．この関数 $l(S, v)$ を最小化することを目標に勾配法を用いてパラメータを調節する．

$$l(S, v) = \sum_{i=1}^{n} T[f(s_i, v) - f(s_{i=0}, v)] \tag{3}$$

3　個人を模倣するプレイヤの作成実験

3.1　本実験の評価関数

本実験の評価関数は，盤面の複数の N-tuple のパターンを評価する部分評価関数の評価値の和を，評価関数の評価値としている．これらの部分評価関数のパラメータを調節することで個人を模倣するプレイヤの作成を目指す．

3.2　パラメータの調整

事前に強化学習を用いたものを初期値にし棋譜学習を行う方法と，初期値のまま棋譜学習を行う方法，それぞれの方法でパラメータの調整を行った．また，TD 学習では 2 つの手法を用意した．TD 学習手法 1 では，棋譜から学習を行う際に石の差を報酬とした．TD 学習手法 2 では，手法 1 に加えて棋譜の手に報酬を与えて学習した．

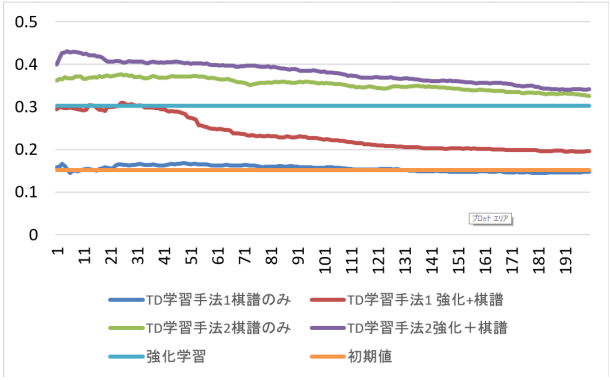

図 1　棋譜との一致率

3.3　実験結果

実験の結果を図 1 に示す．TD 学習の手法 1 では良い結果が得られなかった．TD 学習の手法 2 では，初めは一致率が他の手法よりも高い結果となったが，回数を上げるごとに一致率が下がった．どちらの手法でも強化学習を行った物の方が一致率が高かった．ボナンザメソッドはパラメータの調整が上手くいかなかった．また，どの手法を用いた場合でも回数を増やすごとにテストプレイヤとの勝率が下がった．

4　まとめ

一致率の増加に関しては，手法 2 の方法の強化学習＋棋譜学習が一番良い結果になった．また、予め学習を行った方がしていない方と比べて一致率が高くなった．これは既存の研究と同じ結果である [1]。全体の傾向として、学習を行うごとに勝率が悪くなってしまうことが分かった．原因として，同じ棋譜を繰り返し学習させたため，特定のパラメータが極端に高くなったり低くなっていることが原因と考えられる．これらの結果から，棋譜の手の評価をよくする事で着手の一致率の増加が期待できるが，特定のパラメータのみを増加させることは，プレイヤの強さを下げてしまうと考えられる．特定パラメータのみの増加を防ぎつつ，パラメータを調整する工夫が必要だと考えられる．

参考文献

[1] 山本 智晴, 鶴岡 慶雅，” 将棋における個人に適応した着手推定モデルの構築”, ゲームプログラミングワークショップ 2016 論文集, pp. 112–118, 2016.